INTRODUCTION A L'ANALYSE DES DONNÉES *(Suite)*

« ANALYSE DES CORRESPONDANCES, VALIDITÉ DES RÉSULTATS »

par

L. LEBART

SOMMAIRE

3. ANALYSE FACTORIELLE DES CORRESPONDANCES

3.1. LE PROBLÈME POUR L'UTILISATEUR

L'utilisateur se trouve souvent en présence de « gros » tableaux de correspondances $I \times J$ (ou $I \times J \times K$) (« Contingency tables »), où, à un couple $(i, j) \in I \times J$ (ou un triplet $(i, j, k) \in I \times J \times K$), correspond un nombre.

Nous prendrons comme exemple, pour plus de clarté, un tableau de dimensions modestes : le tableau (8×10) qui croise 8 Modes d'hébergement en vacances et 10 catégories socio-professionnelles. A l'intersection de la ligne du mode de séjour « i » et de la colonne de la catégorie socio-professionnelle « j » se trouve donc un nombre $k(i, j)$ d'individus ou de ménages.

De tels tableaux pourraient, d'ailleurs, être étudiés par l'analyse en composantes principales.

Il est plus naturel de chercher une méthode qui tienne compte du **caractère probabiliste** de ce type de données.

Si $k = \sum_{(i,j) \in I \times J} k(i, j)$; $p(i, j) = \frac{k(i, j)}{k}$ est une estimation de probabilité.

$$p(i) = \sum_{j \in J} k(i, j)/k$$

$$p(j) = \sum_{i \in I} k(i, j)/k$$

peuvent être interprétés en termes de lois marginales.

Dans notre exemple, p(i) caractérise l'importance du mode de séjour « i » et $p(j)$ l'importance de la catégorie socio-professionnelle « j ».

Ce qui est intéressant, lorsque l'on compare les modes d'hébergement de deux catégories socio-professionnelles, c'est de confronter la part de chacun des types de séjour dans le total des séjours, et non pas les nombres absolus de séjours. En un mot, il est surtout intéressant de comparer des probabilités conditionnelles.

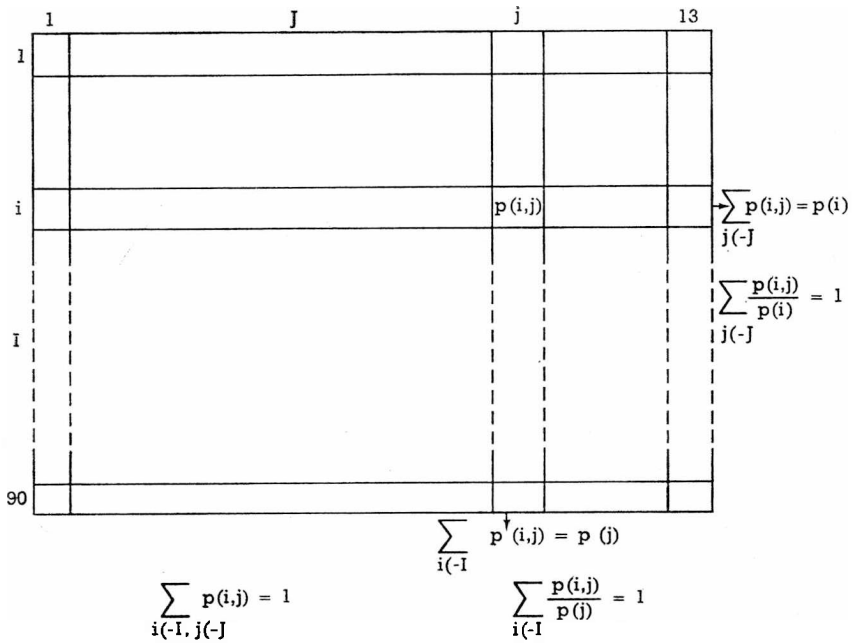
Il s'agit donc de trouver une méthode qui permette de décrire les éventuelles proximités existant entre les lignes et entre les colonnes d'un tableau de correspondance (proximités entre **formes** indépendamment des niveaux ou tailles), compte tenu des poids différents de ces lignes et de ces colonnes.

3.2. LE PROBLÈME POUR LE STATISTICIEN

1) Position du problème

Comme dans le cas des composantes principales, on va se placer dans un espace ayant autant de dimensions qu'il existe d'éléments dans une ligne ou une colonne du tableau de correspondance.

Tableau de correspondance



Choisissons les colonnes pour fixer les idées (pour notre exemple, on se place donc dans un espace à 8 dimensions, dans lequel on aura par conséquent 10 points). Nous verrons plus loin qu'il y a intérêt à prendre comme dimensions celles correspondant au plus petit côté du tableau rectangulaire.

Lorsqu'aucune confusion ne sera à craindre I, J désigneront aussi bien un ensemble que le nombre des éléments de cet ensemble.

Nous n'allons pas, comme nous l'avions fait précédemment, placer directement les valeurs $p(i, j)$ ou $k(i, j)$ sur les axes de cet espace.

Dans l'espace $\mathbf{R}^J$ nous construirons un nuage de I points, chaque point ayant pour coordonnées les quantités :

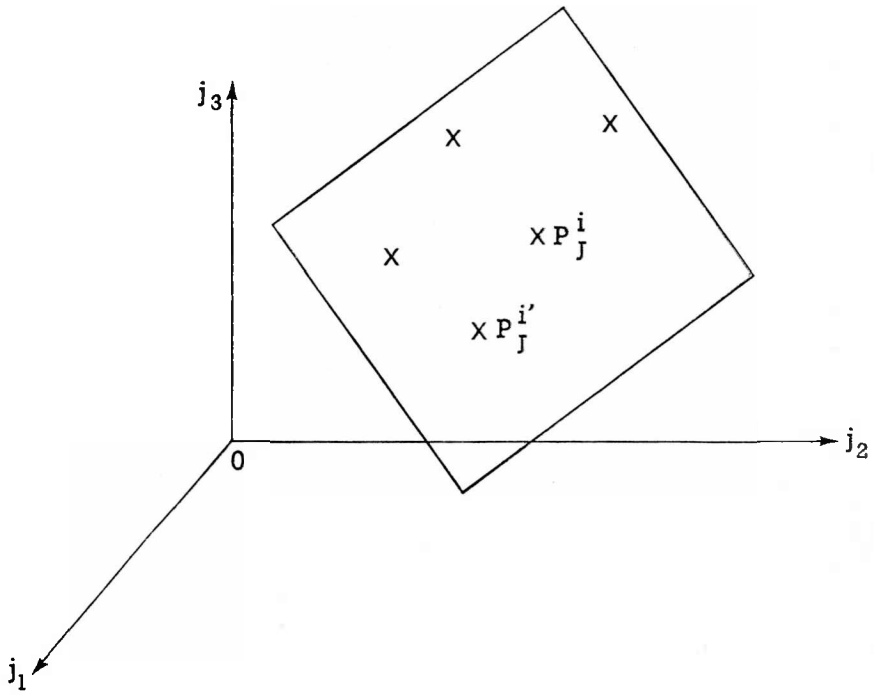
$$\left(\frac{p(i, j)}{p(i)}\right)_{j \in J}$$

et étant affecté de la masse $p(i)$.

Nous noterons p_j^i le vecteur dont les J composantes sont :

$$\left(\frac{p(i, j)}{p(i)}\right)_{j \in J}$$

Figure 2



Remarquons que les I points p_j^i sont tous situés dans un hyperplan (sous-espace à $J-1$ dimensions), puisque leurs coordonnées vérifient la relation :

$$\sum_j \frac{p(i, j)}{p(i)} = 1 \qquad \text{pour } i = 1, 2, \ldots I$$

Remarquons également que si deux points sont proches dans cet espace, cela signifie que les « profils » des lignes représentées par ces points sont voisins.

Il n'est pas raisonnable de mesurer la distance de deux points par la formule classique :

$$(1) \qquad d^2(i, i') = \sum_{j \in J} \left[\frac{p(i, j)}{p(i)} - \frac{p(i', j)}{p(i')} \right]^2$$

En effet, dans cette formule intervient la comparaison terme à terme des J éléments des profils de i et de i', en donnant à ces j éléments le même poids.

Supposons que les effectifs, mesurés par $p(j_0)$, de la colonne j_0 soient considérables.

Dans $d^2(i, i')$ le terme :

$$\left[\frac{p(i, j_0)}{p(i)} - \frac{p(i', j_0)}{p(i')}\right]^2$$

sera très grand par rapport aux autres, et jouera un rôle excessif dans la détermination des proximités.

L'expression pondérée :

$$(2) \qquad d^2(i, i') = \sum_{j \in J} \frac{1}{p(j)} \left[\frac{p(i, j)}{p(i)} - \frac{p(i', j)}{p(i')}\right]^2$$

a le mérite d'atténuer ces disparités.

Elle a également l'avantage de vérifier le principe « d'équivalence distributionnelle », c'est-à-dire que si deux points P_j^i et $P_j^{i'}$ sont confondus et si on les considère comme un seul point affecté de la somme des masses de i et de i', alors les distances entre les éléments de J ne sont pas modifiées.

Cette propriété de la formule (2) est fondamentale. Elle explique la stabilité des résultats issus de ce type d'analyse.

2) **Résolution du problème**

Le problème se situe donc dans le cadre du chapitre 2, puisque la distance entre p_j^i et $p_j^{i'}$ n'est pas une somme de carrés.

Cependant, la métrique Q utilisée ici est particulièrement simple, puisque la matrice associée à la forme quadratique est diagonale, et vérifie la relation [cf. formule (2)] :

$$q_{jj} = \frac{1}{p(j)} \qquad (q_{ij} = 0 \text{ si } i \neq j).$$

Le changement de base effectué précédemment ($y = Cx$ avec C telle que $Q = {}^tCC$) est ici immédiat puisque Q est diagonale. C est également diagonale (**simple changement de l'échelle des axes**), et telle que

$$c_{jj} = \sqrt{q_{jj}} = \frac{1}{\sqrt{p(j)}} \qquad (c_{ij} = 0 \text{ si } i \neq j).$$

On est donc ramené à une analyse simple en prenant comme coordonnées des points du nuage les quantités :

$$\frac{p(i, j)}{p(i)\sqrt{p(j)}}$$

Le nuage est maintenant dans l'hyperplan H d'équation :

$$(2') \qquad \sum_j \sqrt{p(j)}\ x_j = 1$$

Le terme $v_{jj'}$ de la matrice des covariances V s'écrit [en faisant intervenir les poids $p(i)$, $i \in I$].

$$(3)^* \qquad v_{jj'} = \sum_i p(i) \cdot \frac{p(i,j)}{p(i)\sqrt{p(j)}} \cdot \frac{p(i,j')}{p(i)\sqrt{p(j')}} - (\sqrt{p(j)}\sqrt{p(j')})$$

En effet, la j^{ieme} composante du point moyen (centre de gravité) s'écrit :

$$m_j = \sum_i p(i) \cdot \frac{p(i,j)}{p(i)\sqrt{p(j)}} = \frac{p(j)}{\sqrt{p(j)}} = \sqrt{p(j)}.$$

$v_{jj'}$ s'écrit finalement :

$$v_{jj'} = \sum_i \frac{p(i,j)p(i,j')}{p(i)\sqrt{p(j)}\sqrt{p(j')}} - \sqrt{p(j)}\sqrt{p(j')}.$$

Les facteurs cherchés vérifient l'équation :

$$Vu = \lambda u$$

Soit :

$$(4) \qquad \sum_k \sum_i \frac{p(i,j)p(i,k)}{p(i)\sqrt{p(j)}\sqrt{p(k)}} u_k - \sum_k \sqrt{p(j)}\sqrt{p(k)}u_k = \lambda u_j.$$

Il y a J équations de ce type ($j = 1, \ldots, J$).

Notons que le vecteur propre u^* défini par $u^*_k = \sqrt{p(k)}$ est racine évidente de ce système.

En effet, le premier membre de (4) se réduit à :

$$\sqrt{p(j)} - \sqrt{p(j)} = 0$$

et le second membre à $\lambda\sqrt{p(j)}$.

u^* est donc vecteur propre relatif à la valeur propre 0.

Remarquons que tout vecteur propre u solution de (4), et différent de u^*, est solution de l'équation simplifiée (4') :

$$(4') \qquad \sum_k \sum_i \frac{p(i,j)p(i,k)}{p(i)\sqrt{p(i)}\sqrt{p(k)}} u_k = \lambda u_j.$$

En effet, $u \in H$, support du nuage, défini par l'équation (2') ; par suite :

$$\sqrt{p(j)} \sum_k \sqrt{p(k)}u_k = 0.$$

(La métrique naturelle nous permet d'identifier ici facteurs et axes factoriels et donc de considérer que $u \in H$; la relation ci-dessus peut cependant être établie en remarquant que u est orthogonal à u^*, puisque ce sont des vecteurs propres de la matrice symétrique v).

(*) La relation (3) utilise la décomposition classique de la covariance

$$\sum f_i(x_i - \bar{x})(y_i - \bar{y}) = \sum f_i x_i y_i - \bar{x}\bar{y}$$

Finalement les calculs se résument à ceci :

On cherche les vecteurs propres de la matrice **symétrique** S de terme général :

$$s_{jk} = \sum_i \frac{p(i,j)p(i,k)}{p(i)\sqrt{p(j)p(k)}}$$

La projection du point « i » de l'ensemble I sur le r^{ieme} axe factoriel est égale à $f_r(i)$ tel que :

$$f_r(i) = \sum_k u_{rk} \frac{p(i,k)}{p(i)\sqrt{p(k)}}$$

Comme nous cherchons des facteurs qui s'appliquent sur les données initiales :

$$\frac{p(i,k)}{p(i)} \qquad \text{(sans modification d'échelle)}$$

il nous faudra prendre comme facteur v_r :

$$v_{rk} = u_{rk}/\sqrt{p(k)}$$

N.B. Le vecteur u^* ayant pour coordonnées $\sqrt{p(k)}$, vecteur propre de V correspondant à la valeur propre 0, est vecteur propre de S, correspondant à la valeur propre 1.

Ce vecteur, q-orthogonal à H, n'est pas à prendre en compte dans l'analyse.

3) **Liaisons entre les représentations des ensembles I et J**

Nous venons de représenter les proximités entre les éléments de l'ensemble I, vis-à-vis de leurs associations avec ceux de l'ensemble J.

On pourrait, en renversant les rôles de I et de J dans les calculs ci-dessus, obtenir une représentation analogue des éléments de J. Mais I peut être très grand ($I = 200$ par exemple) ; les calculs faisant alors intervenir des extractions de valeurs propres, relatives à des matrices (I, I), deviennent alors coûteux, ou parfois impossibles. Fort heureusement, il existe des relations simples entre les facteurs représentant I et ceux représentant J :

Les facteurs $u_k/\sqrt{p(k)} = v_k$, issus de l'équation (4'), vérifient :

$$(5) \qquad \sum_k \sum_i \frac{p(i,j)p(i,k)}{p(i)p(j)} v_k = \lambda v_j$$

[(5) est obtenu à partir de (4') en remplaçant u_k et u_j par ses valeurs en fonction de v_k et v_j.]

Sommons les deux membres de l'équation (5) par rapport à j, en pondérant les termes par la quantité $\frac{p(e,j)}{p(e)}$.

On obtient, en intervertissant les signes Σ :

$$(6) \qquad \sum_i \left\{ \sum_j \frac{p(i,j) \cdot p(e,j)}{p(e)p(j)} \right\} \sum_k \frac{p(i,k)}{p(i)} v_k = \lambda \sum_j \frac{p(e,j)}{p(e)} v_j$$

On reconnaît, entre les accolades, le terme général d'une matrice analogue à celle de l'équation (5), où i et j, k, deviennent maintenant respectivement j, e, i.

Posons :

$$\sum_k \frac{p(i,k)}{p(i)} v_k = w_i'.$$

Alors l'équation (6) nous montre que le vecteur w', dont les composantes sont les w_i', est un vecteur propre qui joue pour J le même rôle que le vecteur v pour I.

On passe donc très simplement de la représentation de l'ensemble I à celle de J.

Un point reste cependant à éclaircir, celui de la norme du vecteur w'.

Calculons-la :

$$\sum_i p(i) w_i'^2 = \sum_i \left(\sum_k \frac{p(i,k)}{p(i)} v_k \right)^2 \cdot p(i)$$

$$= \sum_i \sum_k \sum_{k'} \frac{p(i,k)p(i,k')}{p(i)} v_k v_{k'}.$$

D'après l'équation (5) :

$$= \sum_{k'} (\lambda v_k'^2 \cdot p(k')) = \lambda \sum_{k'} p(k') v_k^{2\prime}$$

$$= \lambda \text{ (puisque } V \text{ est } q. \text{ normé).}$$

Posons maintenant :

$$(7) \qquad w_i = \frac{1}{\sqrt{\lambda}} \sum_k \frac{p(i,k)}{p(i)} v_k.$$

Le vecteur ainsi obtenu est donc de norme 1.

Ainsi, les facteurs du nuage J sont proportionnels aux coordonnées des points représentatifs de I sur les axes factoriels du nuage I (et réciproquement).

3.3. REPRÉSENTATION SIMULTANÉE DES ENSEMBLES I ET J

On peut représenter les proximités entre les éléments de I dans le plan des deux premiers axes factoriels (les coordonnées du point « i » sont alors les nombres w_{1i} et w_{2i}).

On peut représenter sur le **même** graphique les proximités entre les éléments de J (les coordonnées du point « j » étant les nombres v_{1j} et v_{2j}). Cette représentation simultanée est justifiée par le fait que les w_i apparaissent comme des barycentres des v_j, chaque v_j étant affecté du poids $\frac{p(i, j)}{p(i)}$, probabilité conditionnelle d'occurrence de « j » sachant que « i » est réalisé.

Dans notre exemple, sur lequel on pourra trouver plus de détails plus bas, le mode d'hébergement « i » est d'autant plus près de la catégorie socio-professionnelle « j » que celle-ci intervient fortement dans le profil de ce mode d'hébergement.

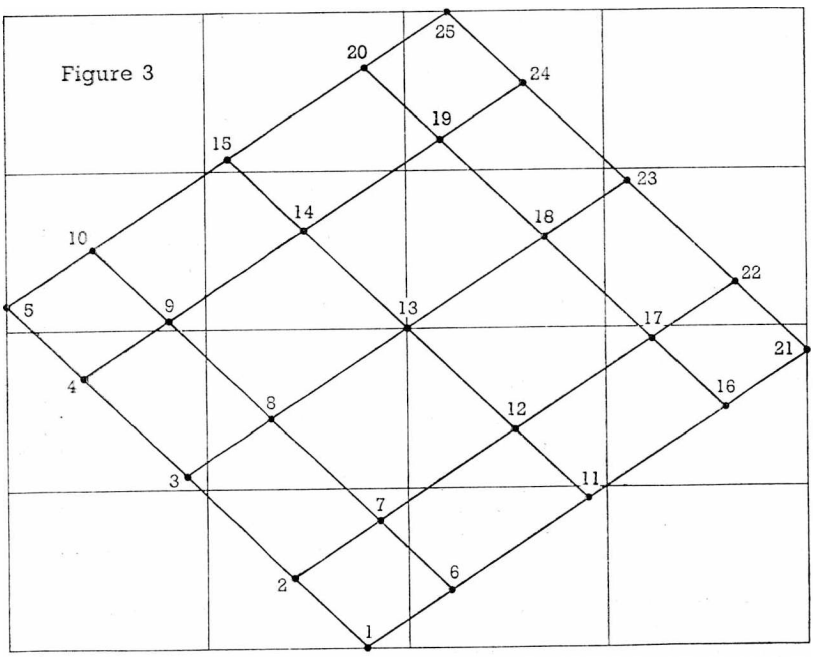
3.4. EXEMPLES D'APPLICATION

L'analyse des correspondances possède un bon pouvoir descriptif des tableaux de nombres positifs. (Sans que ceux-ci soient forcément des tableaux de probabilités ou de fréquences).

Les résultats obtenus lorsqu'interviennent des variables qualitatives (variables prenant les valeurs 0 ou 1) sont sensiblement meilleurs que ceux que donne l'analyse en composante principale.

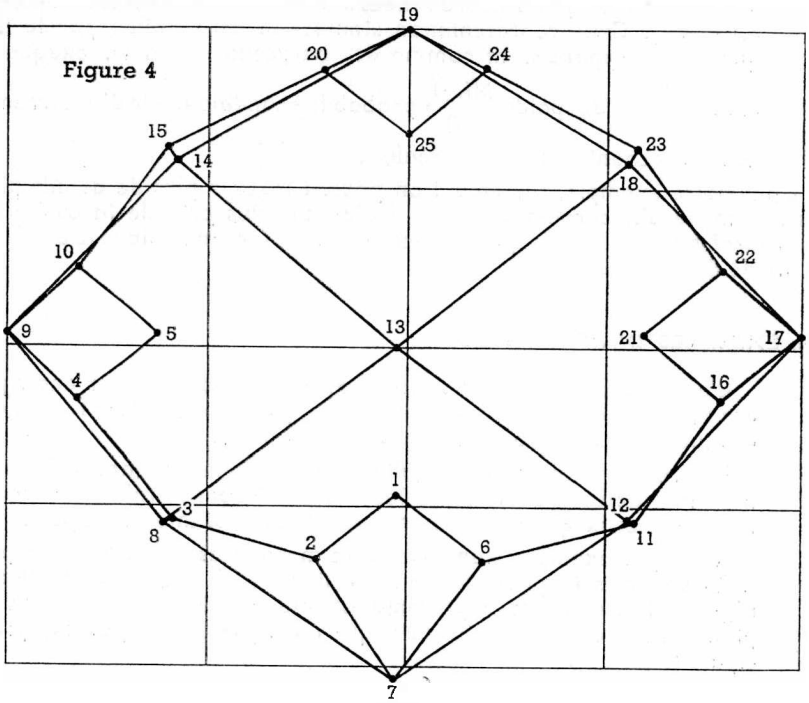
Ainsi, par exemple. l'analyse de la matrice associée à un graphe à 25 sommets (représentant un damier 5×5) par cette méthode (figure 3) donne une représentation plus « satisfaisante » dans le plan des deux premiers facteurs que l'analyse en composante principale (figure 4).

Figure 3



(COMMENTAIRES DANS LE TEXTE)

Figure 4



Application au tableau croisé (8 × 10) (1) : « Modes d'hébergement en vacances. Catégories socio-Professionnelles »

TABLEAU I

Modes d'hébergement en vancances

Catégorie socio-professionnelle du chef de ménage :	Nombre (en milliers)	Répartition (en pourcentage)								
		Ensemble	Hôtel, pension de famille	Maison louée, chez l'habitant	Maison en propriété	Chez des parents	Chez des amis	Tente ou caravane	Village de vacances	Divers (1)
Agriculteurs	796	100	20,1	3,5	—	40,4	4,5	17,7	5,7	8,1
Salariés agricoles	260	100	13,4	13,1	0,4	68,5	3,1	—	1,5	—
Patrons de l'industrie et du commerce	2 978	100	23,5	11,9	7,7	32,2	6,2	9,8	4,0	4,7
Cadres supérieurs et professions libérales	4 620	100	20,8	10,2	13,7	34,2	6,6	7,8	3,5	3,2
Cadres moyens	4 298	100	13,3	12,5	6,5	39,3	4,8	17,4	3,6	2,6
Employés	2 972	100	14,8	13,6	5,6	36,3	6,0	14,6	6,0	3,1
Ouvriers	9 209	100	8,5	12,1	4,2	44,0	5,4	15,9	5,7	4,2
Personnels de service	583	100	11,1	7,4	3,6	50,4	13,6	9,8	3,1	1,0
Autres actifs	1 423	100	5,4	4,2	13,3	59,0	3,7	8,7	2,0	3,7
Non actifs	3 940	100	18,8	8,4	8,3	45,4	7,9	6,0	2,6	2,6

(1) Gîtes, maisons familiales de vacances, auberges de jeunesse, colonies de vacances, etc...

(1) Cet exemple est extrait d'un travail de J. Carayon, portant sur des données statistiques issues de l'étude de M. GOGUEL, « Les vacances des Français en 1964 », **Études et conjoncture**, juin 1965.

L'analyse factorielle de ce tableau par la méthode précédente fait apparaître un premier facteur représentant 53 % de la dispersion totale, et un second facteur expliquant 24 % de cette dispersion. La figure 5, qui donne les configurations des points dans le plan des deux premiers facteurs, rend donc compte de 77 % de la variance totale du tableau analysé.

Un pourcentage aussi élevé est dû essentiellement à la petite dimension de l'espace de départ (8), et au petit nombre de points (10).

Autrement dit, il n'est pas « trop difficile » de trouver un espace à 2 dimensions qui ajuste approximativement 10 points dans un espace à 7 dimensions (les profils sont, comme nous l'avons vu, dans une variété linéaire à 8 — 1 = 7 dimensions).

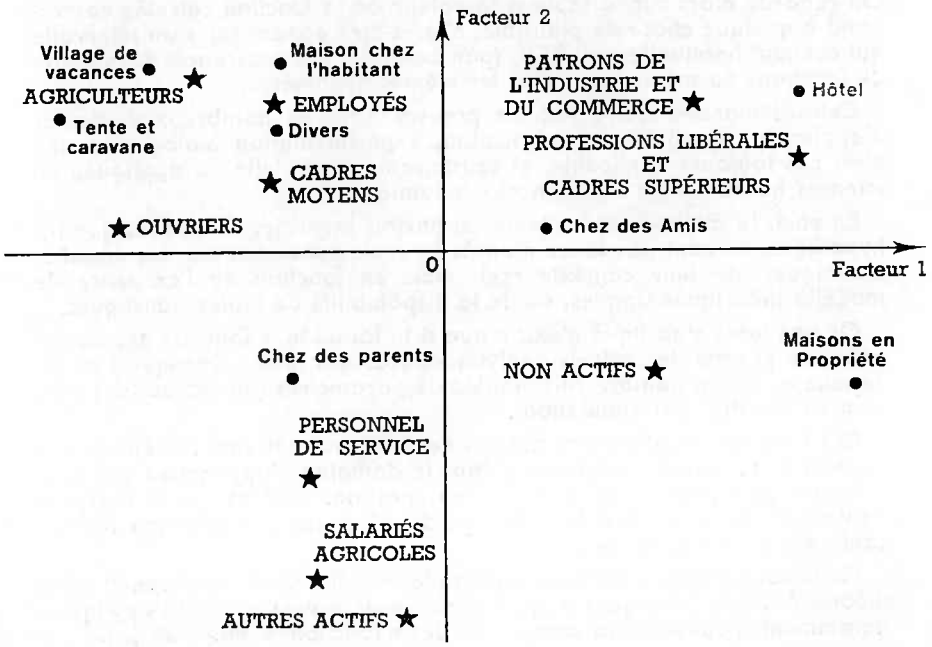
Nous verrons plus précisément au chapitre suivant ce qu'il faut entendre par-là.

La figure 5 donne évidemment une image de la correspondance beaucoup plus vivante et assimilable que le tableau de contingence initial. Son élaboration étant automatique (les positions des points, repérés par trois caractères alphanumériques, sont fournies par l'imprimante de l'ordinateur) elle est donc obtenue sans délais, et constitue un complément utile aux programmes classiques de tris croisés, lorsque le nombre de modalités mises en correspondance est élevé (dépasse 4 ou 5).

On observe sur la figure 5 des proximités entre modes d'hébergement : la proximité de « village de vacances » et de « tente et caravane » traduit le fait que ces deux modes d'hébergement ont mêmes profils socio-professionnels.

Figure 5

Proximités entre modes d'hébergement en vacances et catégories socio-professionnelles



On observe également des proximités entre catégories socio-professionnelles : les « patrons de l'industrie et du commerce » et les « professions libérales et cadres supérieurs » ont des profils d'hébergement voisins.

Enfin, on observe des proximités « croisées », entre modes d'hébergement et catégories socio-professionnelles, qui nous renseignent non plus sur les similitudes de profils, mais sur leur composition : ainsi, les ouvriers prennent souvent leurs vacances en « tente et caravane », très rarement dans un hôtel ou dans une maison leur appartenant, etc.

Nous renvoyons le lecteur à la section 4.3 pour quelques précisions concernant la validité des résultats issus de ce type d'analyse.

4. CONTRÔLE DE VALIDITÉ DES RÉSULTATS EN ANALYSE FACTORIELLE

4.1. TESTS D'HYPOTHÈSE ET SIMULATION

a) La validité des résultats en statistique

Les méthodes d'analyse factorielle dont nous avons parlé jusqu'à présent ont un assez grave inconvénient : elles fournissent toujours un résultat ! Il s'agit d'un inconvénient malheureusement familier en statistique : un simple calcul de moyenne, ou de régression, fournit également toujours un résultat, considéré généralement comme l'estimation d'un paramètre idéal, dont l'existence découle d'hypothèses concernant la population théorique qui est supposée avoir généré les observations.

Les aspects pratiques de la démarche du statisticien sont généralement les suivants : une hypothèse, qui peut être simple ou composite, est faite au sujet de la population parente. Nous désignerons cette hypothèse par H_0.

Une certaine fonction des observations est alors calculée, dont on sait que, sous l'hypothèse H_0, elle suit une certaine loi, préalablement tabulée.

On regarde alors sur la table si la valeur de la fonction calculée correspond à quelque chose de plausible, c'est-à-dire appartient à un intervalle qui contient habituellement 95 % (par exemple) des occurences des valeurs de fonctions du même type, sous les mêmes hypothèses.

Cette démarche, qui a fait ses preuves dans de nombreux domaines d'application (contrôle de fabrication, expérimentation biologique, etc.) n'est pas toujours applicable, ni satisfaisante lorsqu'elle est appliquée en sciences humaines ou en sciences économiques.

En effet, la méthode précédente est parfois inversée, en ce sens que les hypothèses ne sont pas faites d'après la seule considération des données statistiques, de leur contexte réel, mais en fonction de l'existence de modèles théoriques simples, ou de la disponibilité de tables statistiques.

Or une table statistique n'existe que si la loi de la « fonction des observations » permet des calculs analytiques (cas des tables classiques) ou ne dépend que d'un nombre raisonnable de paramètres (au cas ou les tables seraient établies par simulation).

Ces fonctions relativement simples des observations sont forcément très limitées et rarement adéquates : dans le domaine d'application qui nous intéresse plus particulièrement, les observations sont rarement indépendantes, ou ont rarement le même poids, d'où une complication inextricable des éventuels modèles...

Nous nous trouvons bien, en analyse factorielle, dans une situation où la théorie des tests classiques ne peut pleinement et valablement s'appliquer, notamment à cause de la complexité des « fonctions » mises en jeu.

Le problème essentiel est de savoir ce que valent les représentations obtenues dans l'espace des premiers facteurs : il nous faut donc connaître la loi des valeurs propres calculées au cours des analyses, afin de savoir si elles sont « anormalement » élevées, et donc si les facteurs qui leur correspondent extraient bien une part significative de la dispersion totale. Connaître la loi des valeurs propres permet donc de savoir quelle est la dimension « n » du sous-espace de représentation, formée par conséquent des « n » premiers facteurs.

Mais ici, d'une part l'hypothèse H_0 sous laquelle on peut calculer la loi des valeurs propres est souvent beaucoup trop restrictive (variables normales indépendantes d'écart-type unité, par exemple), d'autre part, même sous une telle hypothèse, la complexité des résultats obtenus les rend très difficilement utilisables.

Pour fixer les idées, si les variables analysées sont normales, multidimensionnelles, de matrice des covariances théoriques égale à la matrice unité, la densité de probabilité des valeurs propres s'écrit, en fonction du nombre n d'observations et du nombre p de variables :

$$dF = \frac{\pi^{\frac{P}{2}}}{2^{\frac{p(n-1)}{2}}} \prod_{j=1}^{p} \frac{\lambda_j \frac{1}{2}^{(n-p-2)} \exp\left\{-\frac{1}{2}\Sigma\lambda_j\right\}}{\Gamma\left(\frac{1}{2}(n-j)\right)\Gamma\left(\frac{1}{2}(p+1-j)\right)} \prod_{j<k}(\lambda_j - \lambda_k) \prod_{j<k} d\lambda_j$$

Cette formule est pratiquement inutilisable ; il faut donc procéder autrement pour savoir combien de facteurs retenir lors des analyses factorielles.

b) **Les programmes-tests**

Supposons que l'on veuille savoir si la première valeur propre peut vraisemblablement provenir d'un échantillon E pour lequel l'hypothèse H_0 est vérifiée. Il nous suffit de générer quelques échantillons simulés E_i, et de calculer les différentes valeurs propres $f(E_i)$ que l'on comparera à la valeur initiale $f(E)$.

Si $n - 1$ réalisations ont été simulées, et si la loi de $f(E)$ est la même que celle des $f(E_i)$, autrement dit si H_0 est vérifiée, alors $f(E)$ a une chance sur n d'être supérieur aux $f(E_i)$.

Si l'on réalise par exemple 19 simulations E_1, E_2, ... E_{19}, la valeur observée a une chance sur 20 (c'est-à-dire 5 chances sur 100, seuil usuel en statistique) d'être supérieure à l'ensemble des valeurs simulées.

En pratique, pour des analyses de tableaux particulièrement importants, on pourra se limiter à un nombre beaucoup plus restreint de simulations, en instituant des procédures d'arrêt : (ceci afin d'éviter une exploitation qui peut être coûteuse en temps-machine).

1) Arrêt si la valeur observée est dépassée par une valeur simulée ;

2) Arrêt si la valeur observée est bien plus grande que, par exemple, les 5 premières valeurs simulées, à l'aide d'une sorte de « t » de Student, ou de toute autre fonction permettant d'apprécier la distance d'une observation à un petit échantillon.

Il reste donc, pour chaque type de problème, à choisir une hypothèse H_0 convenable, et à générer des échantillons sous cette hypothèse.

Nous allons rappeler sur quels types de données on est conduit à procéder à des analyses factorielles, en pratique.

Ce sont :

— Les tableaux de valeurs numériques continues.

— Les tableaux de contingence (correspondances).

— Les tableaux de correspondances ensemblistes (dans les cases ne figurent que les valeurs 0 ou 1).

— Les tableaux mixtes (valeurs discrètes ou mélange de valeurs continues et discontinues).

Pour chacun de ces cas, nous allons voir quels sont les hypothèses H_0 les mieux adaptées, et nous préciserons quels sont les types de simulation à effectuer.

4.2. TABLEAUX DE VALEURS NUMÉRIQUES CONTINUES : ANALYSE FACTORIELLE DES RANGS

Ce type de données se présentent habituellement sous la forme de tableaux rectangulaires, dont les lignes par exemple, sont des variables, les colonnes représentant des observations de ces variables.

Une hypothèse H_0 naturelle est l'indépendance des diverses variables.

Cependant, la simulation de nouveaux échantillons sous l'hypothèse H_0 nous oblige à spécifier la forme des distributions de chacune des variables ce qui est embarrassant, car il peut y avoir autant de types de distributions que de variables, ou bien des lois de distributions difficiles à rattacher à un type connu, ou encore la spécification elle-même peut n'avoir aucun sens réel.

Il est alors utile de substituer à l'analyse en composantes principales qui s'impose généralement pour ces tableaux, une analyse de rangs.

La méthode consiste à définir la distance entre deux variables I et I′ par la formule :

$$d^2(I, I') = \frac{6}{n(n+1)(n-1)} \sum_J [R(I, J) - R(I', J)]^2$$

$R(I, J)$ est le rang de la $J^{\text{ième}}$ observation de la $I^{\text{ième}}$ variable lorsque les n observations de cette variable sont classées par ordre de grandeur. On reconnaît en « r » $= 1 - d^2(I, I')$, le coefficient de corrélation des rangs de Spearman, relatif au couple de variables I et I'.

Autrement dit, le tableau des rangs $R(I, J)$ sera substitué dans l'analyse au tableau des valeurs numériques de départ, ce qui présente les avantages suivants :

1) La linéarité des liaisons n'est plus privilégiée puisque la distance $d(I, I')$ est invariante par toute transformation monotone croissante des variables.

2) Les résultats obtenus s'interprètent assez aisément en terme de rangs et de classement. (Par exemple, la proximité entre deux points variables dans le plan des deux premiers facteurs signifiera que les classements des observations de ces variables sont similaires, la proximité entre deux points-observation indiquera une similitude des rangs de ces observations pour les diverses variables, notions qui restent claires même lorsque l'ensemble des variables étudiées est très hétérogène).

3) Il est maintenant facile de tester la significativité des valeurs propres, car le tableau de rang $R(I, J)$, dans l'hypothèse d'indépendance des variables, est indépendant de la forme des distributions du tableau de

départ : il suffira donc de prendre une table de nombres au hasard (directement fournie par une fonction-bibliothèque ou un sous-programme) et de classer les éléments des lignes, pour obtenir un échantillon vérifiant H_0.

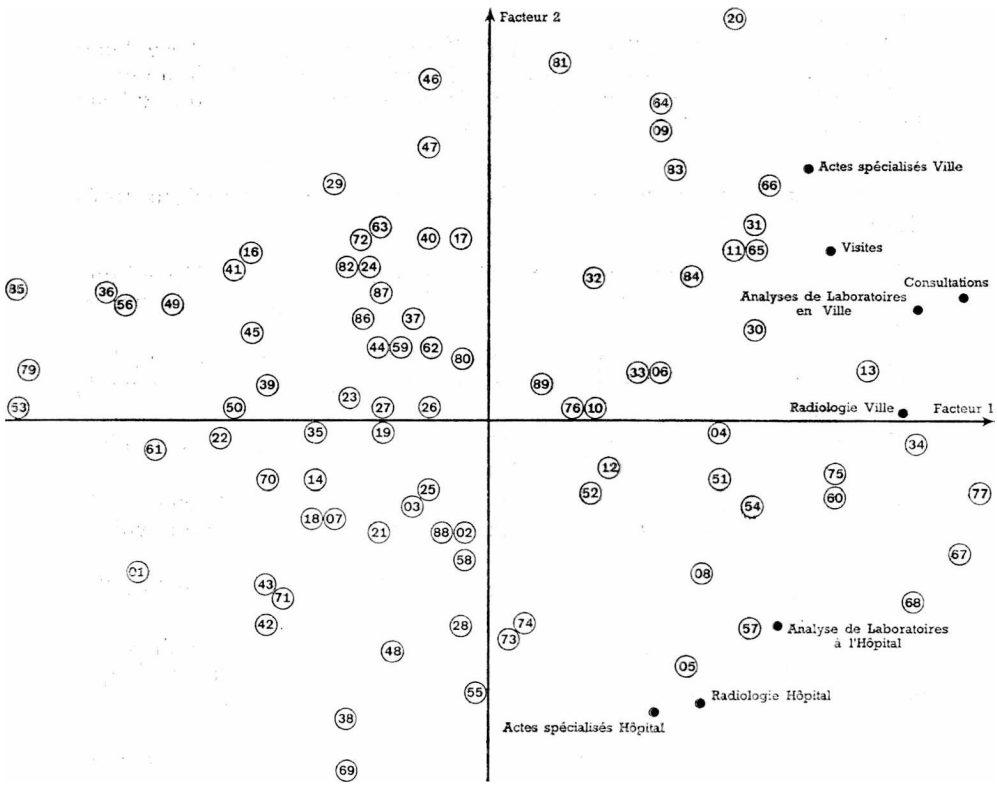
Exemple d'application de l'analyse des rangs

Étude des divers classements des départements français vis-à-vis de 8 postes de leurs consommations médicales :

Nous allons utiliser de nouveau les variables déjà mentionnées à propos de l'analyse en composantes principales : ce sont les consommations, rapportées au nombre de personnes protégées du régime général de la Sécurité Sociale, d'analyses de laboratoire, d'actes de radiologie, d'actes de spécialistes (actes cotés en K) (pour ces trois rubriques, on distinguera actes dispensés en ville et à l'hôpital), de consultations, de visites.

Le tableau de données initiales est donc converti en tableau de rangs, dont l'analyse conduit à la représentation de la figure 6 (1), dans le plan des 2 premiers facteurs.

Figure 6



(1) Les départements sont désignés par leurs numéros minéralogiques.

Les résultats s'interprètent aussi aisément que ceux de l'analyse en composantes principales : toutes les variables sont situées à la droite du graphique, ce qui traduit le fait qu'il n'y a pas de classement inversé. Ainsi, les départements situés sur la gauche du graphique sont en queue de classement simultanément pour toutes les variables. Il y a par contre une opposition entre le haut et le bas du graphique, qui traduit une divergence entre les classements pour les consommations de villes, et ceux relatifs aux consommations d'hôpital. En bas du graphique se trouvent des départements ayant un bon rang pour les consommations hospitalières.

Sans insister sur l'interprétation de ces résultats, voyons ce que l'on peut dire de la validité de cette représentation.

Les deux premiers facteurs de l'analyse représentent respectivement 44 et 25 % de la trace, c'est-à-dire de la variance totale, le troisième facteur 10 % seulement.

19 tableaux de même dimension (8 × 90), ont été simulés. Pas une seule fois, les valeurs 44 et 25 % n'ont été dépassées par le premier et le second facteur. Par contre, la valeur 10 % a toujours été dépassée par les troisièmes facteurs de ces analyses.

Il est donc clair que les deux premiers facteurs extraient une part anormalement forte de la dispersion, tandis que le troisième n'est pas à prendre en compte, puisque l'analyse d'un tableau où les rangs proviennent de permutations aléatoires fournit un troisième facteur d'une variance plus forte.

Nous considérons en effet le domaine d'occurence habituelle des valeurs propres issues de l'analyse d'un tableau des rangs d'une table de nombres au hasard comme un « bruit de fond ». Toute valeur inférieure à ce « bruit de fond » est donc réputée non significative.

Résultats simulés concernant les tableaux de dimensions inférieures à 25 × 75

Puisque la loi des valeurs propres d'un tableau de rang ne dépend que de 2 paramètres (longueur et largeur du tableau), il est possible de procéder à une tabulation.

Les trois graphiques suivants correspondent respectivement aux longueurs 25, 50, 75, et, sur chacun d'eux sont portés en abcisse les largeurs 5, 10, 15, 20.

En ordonnées figurent les pourcentages moyens d'inertie dont rendent compte les trois premiers facteurs des analyses de rangs.

Près de chacun de ces points moyens ont été notées les bornes de la zone ayant approximativement 90 chances sur 100 de contenir une valeur propre particulière.

Chaque point des graphiques correspond à 100 analyses factorielles simulées.

Des tables plus complètes et plus précises pourraient être dressées, mais l'utilisation d'un programme-test, adapté aux dimensions de chaque problème particulier, diminue considérablement leur intérêt.

4.3. VALIDITÉ DES RÉSULTATS EN ANALYSE DES CORRESPONDANCES

Nous distinguerons les tableaux de contingence [où dans la case (i, j) figure un effectif $k(i, j)$], les tableaux de correspondance ensemblistes (formés de 0 et de 1), qui peuvent être considérés, lorsque les ensembles I

Figure 7

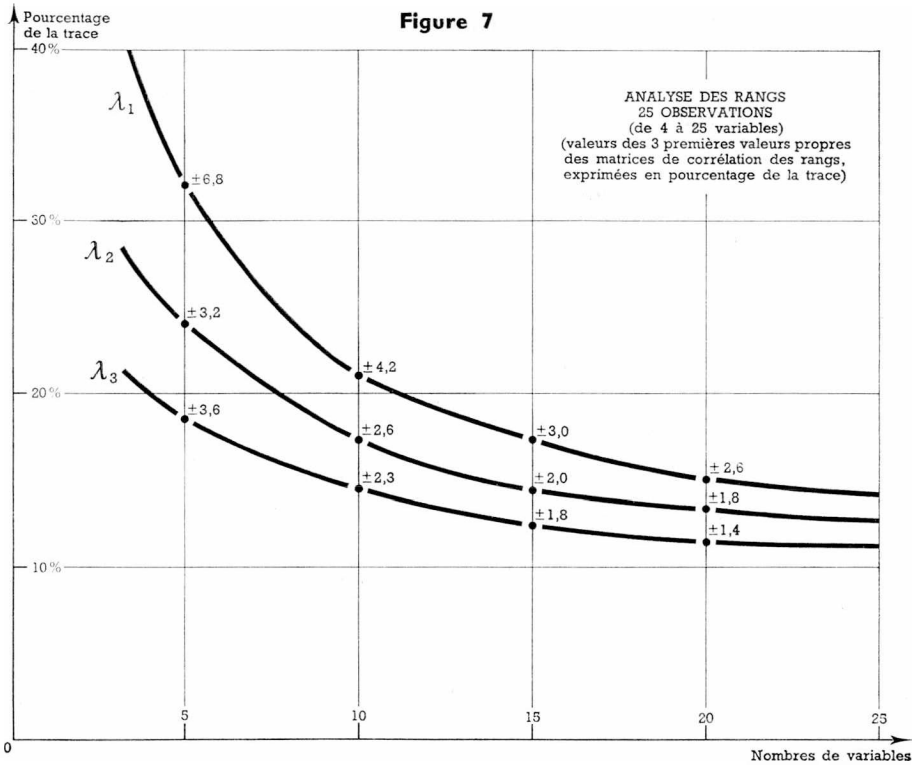


Figure 8

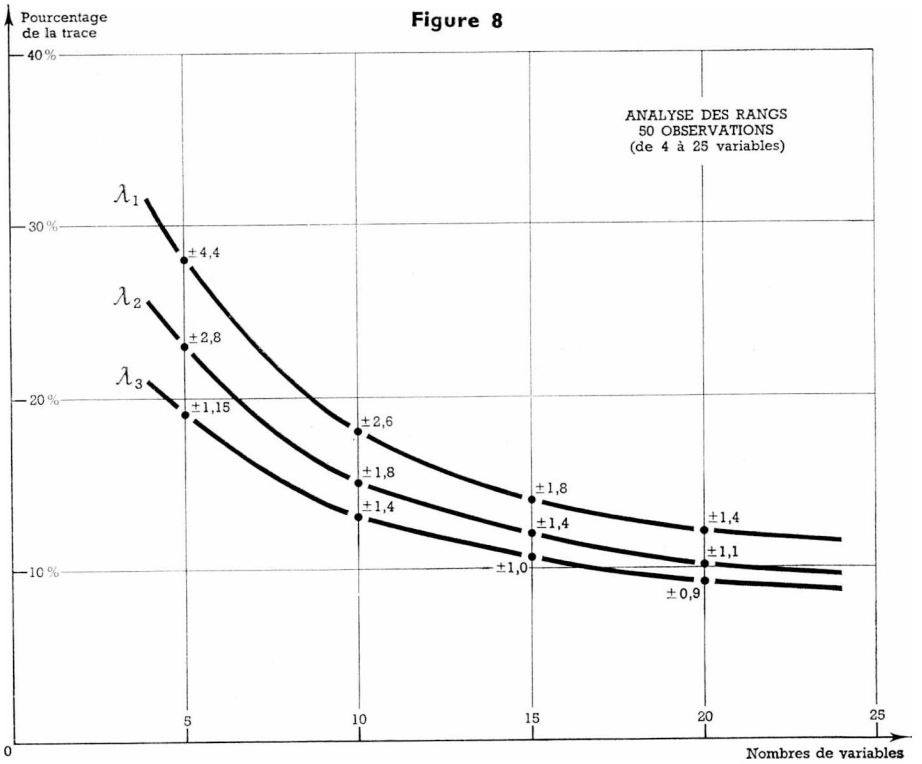
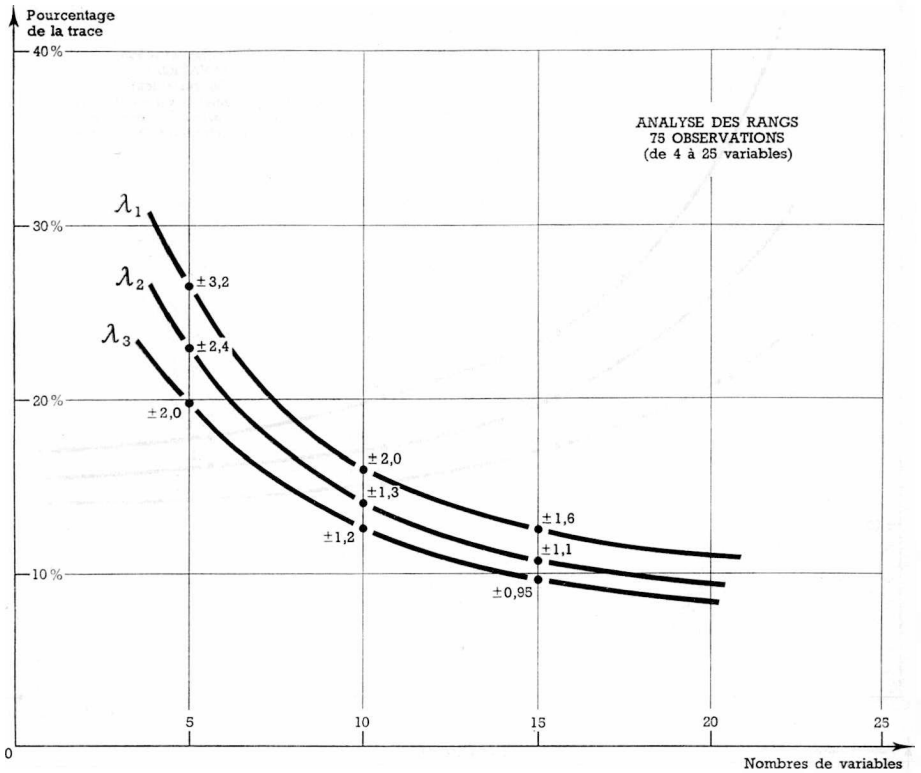


Figure 9



et J coïncident, comme des matrices associées à un graphe ; enfin, les tableaux mixtes, que constituent très fréquemment les résultats d'enquêtes psycho-sociologiques (observations de variables discrètes ayant plusieurs valeurs, par exemple notes de 1 à 5, etc.).

4.3.1. Tableaux de contingence

L'hypothèse H_0 qui s'impose le plus souvent pour ce type de données est la suivante : le tableau observé $P(i, j)$ peut-il être considéré comme issu d'un tableau théorique $P_0(i, j)$ qui ne serait que le produit de ses marges : $P_0(i, j) = P_0(i)P_0(j)$.

Autrement dit, est-ce que les diverses modalités mises en correspondance sont indépendantes, les associations observées en pratique sont-elles significatives ou le reflet de fluctuations d'échantillonnage ?

a) Test global

Il existe pour tester globalement cette hypothèse un test classique : la comparaison de la quantité :

$$S = k \sum_i \sum_j \frac{[P(i, j) - P(i)P(j)]^2}{P(i)P(j)}$$

à un χ^2 à $(I - 1)(J - 1)$ degrés de liberté (k est l'effectif total).

Cette quantité peut d'ailleurs s'écrire, en développant le carré du numérateur :

$$S = k\left(\sum_i \sum_j \frac{P(i,j)^2}{P(i)P(j)} - 1\right)$$

Par comparaison avec la trace T de la matrice dont on extrait les valeurs propres (cf. p. 71), on peut écrire :

$$S = k(T - 1)$$

Ainsi, la somme des valeurs propres correspondant à des facteurs non triviaux est proportionnelle à un χ^2 à $(I - 1)(J - 1)$ degrés de liberté.

Ceci nous donne donc un test global de l'association entre les différentes modalités mises en correspondance, mais ne nous donne aucune information sur le caractère « représentatif » des premiers facteurs de l'analyse.

b) **Test relatif à certains points**

La distance d'un point particulier à l'origine des axes factoriels est, également dans l'hypothèse d'indépendance H_0, proportionnelle à un χ^2, et sa projection dans l'espace des r premiers facteurs peut être comparée, à un coefficient près, à un χ^2 à r degrés de liberté, pourvu que ce point n'ait pas joué un rôle trop important dans la détermination de ce plan ; ceci suppose que sa masse soit faible, ou encore que le nombre de points soit élevé.

En pratique, si un point « j » correspond à un effectif $k(j)$, on comparera sa distance à l'origine dans le plan des 2 premiers facteurs à la quantité $t(j)$:

$$t(j) = \sqrt{5{,}99/k(j)}$$

(La valeur 5,99 correspond à celle qu'un χ^2 à 2 degrés de liberté ne dépasse que dans 5 % des cas).

On peut également tracer dans ce plan des 2 premiers facteurs des cercles centrés en chaque point « j », de rayon $t(j)$: ceux d'entre eux qui contiennent l'origine indiquent alors que leur centre est un point suspect, dont la position ne doit pas être interprétée sans précaution. Cette représentation permet d'apprécier rapidement la confiance que l'on peut accorder à chaque point.

c) **Programmes-tests pour l'analyse des tableaux de contingence**

Une correspondance entre deux ensembles I et J, si elle vérifie l'hypothèse H_0 précédente, dépend tout de même de $I + J$ paramètres (rappelons que les lettres majuscules désignent aussi bien les ensembles que leurs cardinaux).

Il est impossible ici de procéder à une tabulation, même empirique.

Il nous faudra donc générer des tableaux de contingence ayant les mêmes marges que le tableau de données initial, mais qui vérifient l'hypothèse H_0 précédente. Une simulation de ce type se fait selon le schéma multinomial suivant :

Chaque case (i, j) du tableau est affectée de la probabilité théorique $P(i)P(j)$; si k désigne l'effectif total du tableau observé, on répartira k individus fictifs entre les $I \times J$ cases, en tenant compte des probabilités précédentes.

On peut distinguer deux types de programmes :

α) Si k est faible, nous obtiendrons une variable $S(i, j)$ multinomiale par la procédure suivante :

Le tableau cherché $S(i, j)$ est représenté par une zone mémoire mono-indicée $T(L)$, l'indice L pouvant donc varier de 1 à IXJ; (à chaque case (i, j) correspond une valeur bien déterminée de l'indice L, par exemple, $L = I\,(j - 1) + i$).

Un tableau $k\,(L)$ indexé de la même façon donne, pour chaque valeur de (i, j), donc de L, l'effectif « théorique » $k(L)$ qui est l'entier le plus proche de la quantité $kP(i)P(j)$. Au départ, $T(L) = 0$ paur tout L.

Un tableau de travail $M(N)$, occupant k mémoires, est rempli de la façon suivante : l'indice L variant de 1 à $I \times J$, on calcule tout d'abord les effectifs cumulés $K(L)$ (par exemple $K(1) = k(1)$, $K(2) = k(1) + k(2)$, ... $K(I \times J) \simeq k$) et on fait $M(N) = L$, pour tout N compris entre $K(L - 1)$ et $K(L)$ (on aura posé $K(0) = 1$).

Ensuite, à chaque tirage d'un nombre uniformément réparti entre 0 et 1, R, on calcule la valeur entière E de $k \times R$. La quantité $M(E) = L_0$ est l'adresse de la case du tableau T à laquelle il convient d'ajouter 1.

L'opération est à recommencer k fois.

β) Si k est grand, on utilise l'approximation normale de la loi multinomiale.

La variable R issue du sous-programme ou de la fonction-bibliothèque est convertie en variable normale de moyenne $kP(i)P(j)$ et d'écart-type $\sqrt{kP(i)P(j)[1 - P(i)P(j)]}$. La quantité obtenue est alors affectée à la case (i, j).

d) **Application à l'exemple du tableau (8 × 10) précédent**

Les effectifs qui figurent sur le tableau 1, page 74, sont l'extrapolation à toute la population de résultats d'enquête. Nous supposerons, ce qui semble assez près de la réalité, que ce tableau a été établi à partir d'un effectif total de 6 000 ménages. Nous allons donc simuler des tableaux ayant des marges voisines, et un effectif total identique, par la méthode indiquée en *b*).

Rappelons que les 2 premières valeurs propres trouvées lors de l'analyse du tableau initial représentaient respectivement 53 et 24 % de la trace.

Pour les tableaux simulés, dont l'inertie totale est beaucoup plus faible, la première valeur propre n'a jamais dépassé, pour 19 essais, le chiffre de 52 %. Par contre, la seconde a trois fois dépassé 24 %.

« Cependant, une comparaison directe entre valeurs non exprimées en pourcentages montre qu'elles sont toutes significatives. »

On peut être tenté de ne retenir qu'un seul axe descriptif, ce qui conduirait à la représentation de la figure 10. Effectivement, cette figure semble bien contenir l'essentiel de l'information concernant cette correspondance. Notons cependant que le choix de l'espace de représentation est ici plus difficile que lors de l'analyse des rangs, où la trace de la matrice diagonalisée était constante.

Les procédures de contrôle par simulation sont heureusement très souples, de part leur caractère empirique, et permettent de moduler les expériences qui conduiront le statisticien et l'utilisateur à acquérir des convictions, et non plus des impressions fugitives.

4.3.2. **Correspondances ensemblistes et tableaux mixtes**

Plusieurs types de simulation seront adaptés à chaque cas particulier.

Prenons le cas d'une correspondance ensembliste symétrique (ou matrice associée à un graphe non orienté). Les représentations par l'analyse factorielle de cette catégorie de données nous fourniront des descriptions planes du graphe. Le modèle de référence le plus adapté pour des simulations semble être celui qui stipule un tirage exhaustif des k arêtes du graphe, toutes les arêtes ayant la même probabilité d'être tirées.

Une procédure de génération de tableaux de ce type est la suivante :

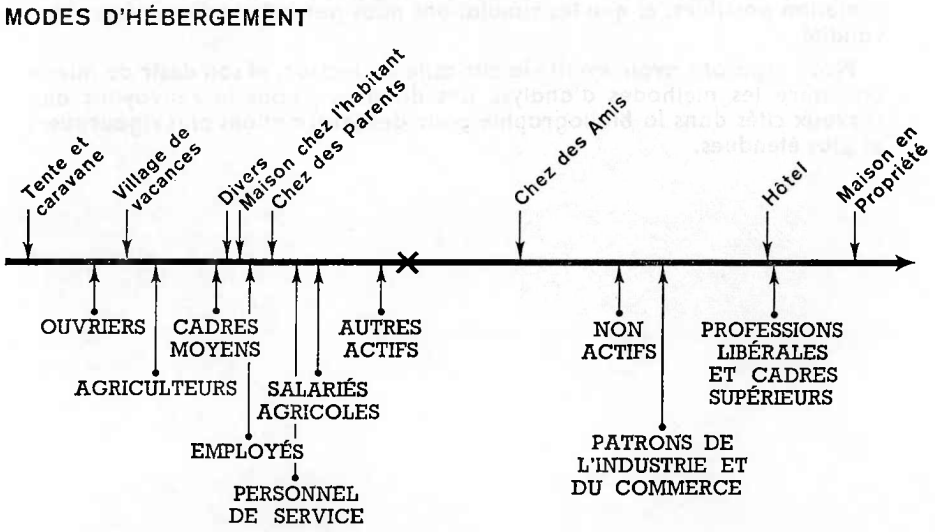
Nous ne nous intéressons qu'à la moitié du tableau de départ, puisque celui-ci est symétrique. Si I désigne la dimension d'un côté de ce tableau, c'est parmi $I \times (I - 1)$ couples possibles que nous devons choisir les k couples qui définissent le graphe. Nous remplissons deux tableaux auxiliaires LAB(J) et LOR(J), de dimensions $I \times (I - 1)$ de façon à appliquer la moitié du tableau (sans la diagonale) sur l'indice J. Chaque case du tableau est repérée par une valeur de J, son « abscisse » est alors LAB(J), et son « ordonnée » LOR(J). En triant une série de nombres au hasard, on génère un arrangement aléatoire de k valeurs possibles de l'indice J. En appelant $K(L, M)$ notre tableau, on fait alors, pour ces k valeurs : $K(\text{LAB}(J), \text{LOR}(J)) = 1$, les autres valeurs étant prises égales à 0. On complète ensuite par symétrie.

Pour les tableaux de valeurs discrètes, l'utilisateur fixera lui-même le schéma de probabilité « a priori » qu'il désire voir tester : par exemple, équiprobabilité des 9 réponses, notées de 1 à 9, à telle question, etc.

Le lecteur désirant avoir des précisions sur ces diverses modalités de calcul pourra consulter un certain nombre de travaux non publiés du **C.R.E.D.O.C.**

Figure 10

Proximités entre mode d'hébergement en vacances et catégories socio-professionnelles (premier facteur de l'analyse)



Conclusion

Transformer de lugubres tableaux de chiffres en paysages statistiques est la préoccupation essentielle de la statistique descriptive ; étendre cette transformation aux données multidimensionnelles, c'est ce que permet l'ordinateur, qui bouscule quelque peu la statistique mathématique classique en passant.

Dans un article intitulé « Idéogrammes : graphiques et géométrie » M. Barbut distingue, pour les figures planes, trois niveaux de propriétés qui sont utilisées pour représenter par le dessin des relations entre objets :

— *Le niveau combinatoire, qui permet de représenter des collections d'objets entre lesquels existent des relations binaires non orientées (par exemple, des graphes, où seules sont significatives des relations d'incidences).*

— *Le niveau de l'ordre, qui permet de représenter des relations d'inclusion, de précédence, des hiérarchies.*

— *Le niveau métrique, où l'égalité et l'addition interviennent, particulièrement adapté à la représentation des structures numériques.*

L'analyse des données, un peu comme le dessin, vise à réaliser un homomorphisme entre les ensembles de valeurs numériques observés, et une représentation visuelle. Elle englobe comme techniques particulières l'analyse factorielle et la classification automatique.

L'analyse factorielle, qu'il s'agisse de l'analyse en composante principale, ou de l'analyse des correspondances, opère au dernier niveau cité ; elle cherche donc une approximation métrique des structures initiales.

La classification automatique, qui fera l'objet d'un prochain article, opère au second niveau, en tentant de mettre en évidence des hiérarchies.

Le problème essentiel est dans tous les cas de savoir si les applications entre structures observées, et structures représentées sont *possibles*, et si elles sont *significatives*.

Quelques éléments de réponses ont été donnés au cours des pages précédentes, puisque l'algèbre nous a fourni quelques techniques de représentation possibles, et que les simulations nous permettent d'apprécier leur validité.

Nous espérons avoir éveillé la curiosité du lecteur, et son désir de mieux connaître les méthodes d'analyse des données ; nous le renvoyons aux travaux cités dans la bibliographie pour des informations plus rigoureuses et plus étendues.

BIBLIOGRAPHIE

Nous rappelons, et nous complétons la liste des ouvrages déjà mentionnés à la fin de la première partie.

L'abréviation L.S.P. signifiera : Laboratoire de statistique de la Faculté des Sciences de Paris, publication multigraphiée.

J. P. BENZECRI, **Distance distributionnelle et métrique du χ^2 en analyse des correspondances,** L.S.P.

J. P. BENZECRI, **Analyse factorielle des correspondances,** L.S.P.

J. P. BENZECRI, **Sur le choix des unités et des poids dans un tableau en vue d'une analyse des correspondances,** L.S.P.

J. P. BENZECRI, **L'approximation stochastique en analyse des correspondances,** L.S.P.

J. P. BENZECRI, **Théorie de l'information et classification d'après un tableau de contingence,** L.S.P.

J. P. BENZECRI, **Sur l'analyse de la correspondance définie par un graphe,** L.S.P.

J. P. BENZECRI, **Leçons sur l'analyse statistique des données multidimensionnelles,** L.S.P.

B. CORDIER (M^me^ Escoffier), « Analyse factorielle des correspondances », (thèse 3^e^ C).

C. R. RAO, **Linear statistical inference and its application,** Wiley and sons, 1968.

A. P. DEMPSTER, **Elements of continuous multivariate analysis,** Addison-Wesley publishing company, 1968.

M. BARBUT, **Idéogrammes, graphiques et géométrie,** Mathématique et sciences humaines, n° 6, 1964.

D. F. MORRISON, **Multivariate statistical methods,** McGraw Hill, 1967.

BIBLIOGRAPHIE

Gaston DESFOSSES, La Bourse des Valeurs (P.U.F., « Que sais-je », 6e éd., 1968).

Jean VALEURS, A quoi sert la Bourse (Seuil, Société, 1re éd., 1966).

S. Robert MILLES, La Grammaire de la Bourse (Flammarion, 2e éd., 1963).

Louis ENGEL, How to buy stocks (Bantam, N. Y., 4e éd., 1967).

NOTE DE LECTURES

Quatre livres pour présenter la Bourse : après les avoir lus, on sait mieux comment d'autres y opèrent, on ne voit pas comment y opérer soi-même. C'est donc sans doute que l'entreprise n'est pas facile ; c'est aussi qu'elle ne s'adresse pas aux épargnants, du moins, à leur porte-monnaie.

Seul Louis Engel écrit dans cette perspective : « le seul livre que vous n'avez pas les moyens de ne pas acheter », au dire de l'éditeur, s'adresse à celui qui n'a pas de valeurs mobilières, pour l'informer de la manière d'en acquérir, et des avantages qu'il peut y trouver. C'est ainsi que non seulement les mécanismes du marché sont présentés de manière vivante (encore que sans doute trop exhaustive, car le détail ici écrase le néophyte), mais qu'encore et surtout les voies d'accès au marché sont amplement décrites : relations avec l'agent de change, rôle des S.I.C.A.V., utilisation des journaux financiers, évaluation des valeurs ; et le détail ici, est utile au néophyte. Ce n'est pas que l'ensemble soit attrayant : c'est sérieux, ce n'est pas publicitaire : quelques chiffres, pas de graphiques, moins encore de photos, L'ouvrage est destiné à celui qui voudrait bien investir, mais ne sait comment le faire ; il n'attire pas celui qui n'a pas ce désir.

Par comparaison, l'ouvrage de M. Robert Milles est décevant. Mal édité, le texte paraît décousu : il est souvent inutile et de plus, non « opérationnel ». Il présente un monde lunaire où des gens parfaitement informés réalisent des opérations de la plus haute complexité, avant de s'en aller « à la campagne, aux eaux ou aux bains de mer ». Ces mêmes opérateurs d'ailleurs se dérangent personnellement pour aller aux guichets du Trésor faire renouveler leurs titres de rente. Banquiers, Agents de Change, journaux, rien de tout cela n'existe. Ce livre est l'œuvre d'un initié s'adressant à d'autres initiés ; et de fait, sa seconde partie, consacrée à la « pratique », est une véritable grammaire des opérations ; utile, peut-être, à l'expert, il ne peut que rebuter l'apprenti.

Trouve-t-on meilleure pâture avec les deux autres ouvrages ? « Le » Défossé est un classique en France (6 éditions depuis 1959). Il est clair, précis, pratique sur certains points : s'il reproduit des pages de la cote en illustration, il ne parle pas non plus des S.I.C.A.V., ni de la manière de passer un ordre. Utile à l'étudiant, il est douteux qu'il réussisse à attirer de nouveaux clients vers le marché.

Il en va de même en ce qui concerne l'ouvrage de Jean Valeurs, pourtant conçu dans une optique différente et originale : plus que les mécanismes, c'est la vie de la Bourse qui est présentée et démystifiée : il s'agit de lever le voile de « secret » qui couvre l'institution aux yeux du public, et par là, de réduire les sombres soupçons que cette ignorance fait peser sur elle. Dans cette perspective, l'auteur a sans doute réussi, mais pas plus que les précédents, il n'a initié à la pratique quotidienne.

Aucun de ces trois manuels ne parvient donc, ni ne cherche vraiment, à attirer le public. Faut-il dès lors s'étonner de constater — comme on l'admet souvent, sans d'ailleurs disposer des moyens de le vérifier — que la clientèle de la Bourse est si étroite, et se renouvelle si peu hors d'un cercle de familles « initiées » ? Jean Valeurs emploie un vocabulaire nouveau, parlant de « marchands d'épargne » et de « produit vendu » ; c'est aussi une perspective nouvelle, qui doit se substituer à celle de la Bourse, « Tiercé du riche ». Mais cette perspective entraîne des exigences d'analyse du marché et d'action promotionnelle. Or, nous avons à peine débuté en ce qui concerne de telles actions.

Les analyses du N.Y.S.E. donnaient les chiffres de 6,5 millions de porteurs en 1952, 8 millions en 1956, 22 millions en 1965. En France, la progression est loin d'avoir été aussi rapide : de 3,5 millions en 1956 à 4 millions environ en 1965, d'après les sources dont on peut disposer. Le retard n'est pas grand, mais la croissance est lente. Ne serait-ce pas dû en partie à une conception anachronique du marché, marché des riches et non marché des masses ? Il est intéressant de relever que les analystes les plus sérieux de la Belle Époque (Neymarck, et aussi Colson lui-même) s'enorgueillissaient de la magnifique diffusion de « nos » valeurs mobilières dans toutes « nos » classes d'épargnants : c'était sans compter avec la concentration des portefeuilles, 95 % des valeurs se trouvant aux mains de 50 % des porteurs : n'y a-t-il pas là à la fois une cause et un effet du peu d'efforts de « commercialisation » des valeurs mobilières ?

P. DHONTE

Les valeurs mobilières en France à la fin du XIX^e^ et au début du XX^e^ siècle (1873-1913) par **Robert GOFFIN** in : Christian Morrisson et Robert Goffin : « **Questions financières aux 18^e^ et 19^e^ siècles** » Travaux et recherches de la Faculté de Droit et des Sciences Économiques de Paris, P.U.F., 1967, 152 p.

Le mémoire de M. Goffin présente avec l'agrément d'une grande clarté, un « bilan » du rôle des valeurs mobilières dans l'économie française de l'époque : bilan qui fait apparaître le paradoxe d'un mécanisme qui, quoique favorable aux épargnants, et très recherché par eux, n'a pas, dans l'ensemble, apporté de contribution satisfaisante à la croissance économique française.

A cet échec, M. Goffin dégage plusieurs raisons ; les unes tiennent aux hommes et à la politique suivie, qui manquent d'esprit d'entreprise ; les autres, aux conditions de fonctionnement du marché, qui sont telles que certains secteurs industriels ne peuvent s'y approvisionner en capitaux frais.

Cet aspect de l'étude mériterait d'être plus approfondi. Mais il est remarquable que l'auteur ait pu disposer d'une bibliographie contemporaine fournie en ce qui concerne les Valeurs Mobilières comme **moyen de placement,** minime en ce qui les concerne comme **moyen de financement.**

Notons enfin que l'auteur parle de cette époque comme étant celle des « petits rentiers ». Cette opinion — courante dans la littérature de l'époque — paraît discutable au regard d'une concentration qui attribue 50 % du portefeuille à 2 % des porteurs, et 80 % du portefeuille à 30 % des porteurs. Plutôt que de la diffusion des Valeurs Mobilières, c'est de leur concentration qu'il faut parler, et c'est dans son analyse qu'il faut rechercher les causes de mauvais fonctionnement du marché.

P. DHONTE

THIN GUYEN HUU, RICHARD (Denise). — Principaux résultats de l'enquête permanente sur la consommation alimentaire des français (données recueillies au cours de l'année 1966).

Études et Conjoncture, n° 10, octobre 1968, pp. 83-154, tabl.

Depuis 1964, chaque année, l'I.N.S.E.E. réalise une enquête sur la consommation alimentaire des Français.

La France, à cet effet, a été divisée en 9 régions ; chaque ménage observé sur une période de 7 jours consécutifs, a dû rendre compte de sa consommation à domicile et hors du domicile en produits alimentaires.

Pour chaque achat de produit ont été relevés : la désignation précise du produit, la dépense correspondante, la quantité correspondante, le lieu d'achat (type du point de vente). Chaque produit alimentaire étant échantillonné pour certains d'entre eux comme le vin, la pâtisserie, la confiserie, il a été difficile d'en connaître les chiffres exacts tant par manque de précision que par refus du renseignement de la part des consommateurs.

En pourcentage, ce sont les achats de viande, volailles, œufs et poissons qui constituent le poste le plus important du budget alimentaire. Ces dépenses représentent 38,4 % de l'ensemble des dépenses, d'alimentation ; le second poste — celui des boissons — ne

représente que 14,3 % Ensuite, viennent les achats de légumes (10,3 %), des produits à base de céréales (10,1 %), du lait et de fromages (9,4 %), de corps gras (7,1 %), de fruits (6 %)...

Cependant les résultats obtenus ont prouvé que l'ampleur des besoins alimentaires ne croissait pas proportionnellement à l'effectif du ménage mais moins rapidement. L'influence de la catégorie socioprofessionnelle est prédominante. Les ménages agricoles, par exemple, vivent en autoconsommation (30 % de leur consommation contre 5 % pour les ménages non agricoles) ; ils disposent d'un moindre choix de produits, par contre ce sont des aliments plus riches en éléments énergétiques.

L'habitat joue aussi. Les ménages qui vivent en ville font une consommation plus importante de produits élaborés : pâtisserie, primeurs, viandes nobles (mouton), plats préparés, fromages, produits surgelés. De façon générale on a noté que le montant du budget alimentaire augmente de manière régulière avec le degré d'urbanisation sauf pour un certain nombre de produits de base à faible élasticité tels que pain, pâtes alimentaires, farine, lait, huile, sucre ou margarine.

Sylvie GUIRAUD

GRANDE-BRETAGNE 1980 par Mark ABRAMS.

Revue « Analyse et Prévision », octobre 1968, n° 4, pp. 665-675.

La situation économique de la Grande-Bretagne en 1960 présente plusieurs traits dont les principaux sont la grande faiblesse de son taux d'accroissement, la prise en charge accrue de l'économie par l'État, le violent contraste existant entre le Nord et le Sud du pays, les barrières entre les classes et entre les races, et l'apparition au sein du secteur privé d'un petit nombre de sociétés géantes. Qu'en sera-t-il en 1980 ? L'évolution dépend de plusieurs facteurs.

I. La population

En 1980, elle sera passée de 55,5 millions à 59,9 millions d'habitants. Cet accroissement pose des problèmes multiples ; pour le logement par exemple, 43 % de la population actuelle vit dans des districts ruraux ou dans des villes ne dépassant pas 50 000 habitants ; ce chiffre atteindra 50 % en 1980 : il faudra donc construire.

Un changement se produira au sein de la composition par âge de la population, seul le nombre des jeunes et des vieux augmentera, alors que le nombre des hommes entre 35 et 64 ans connaîtra une diminution de 4,5 %. Cette évolution implique :

1) Un accroissement des dépenses de santé et d'enseignement de l'ordre de 20 % (il est probable d'ailleurs que l'on demandera aux étudiants et aux malades une participation plus importante aux frais dont ils sont la cause).

2) Un changement dans le monde du travail : la population active n'augmentera pas beaucoup mais cette pénurie quantitative de main-d'œuvre sera compensée sur le plan qualitatif (le nombre de jeunes qui reçoivent une formation post-scolaire a augmenté de 50 % depuis 1960).

3) Des modifications au sein de la famille, 3 types de famille domineront à ce moment-là : sur 19,4 millions de ménages, 25 % seront des ménages vivant seuls, des retraités. 35 % des ménages avec des enfants jeunes, ayant du mal à vivre. 40 % ayant des enfants plus âgés, et des revenus plus importants.

II. Le cadre de la vie économique en 1980

Par rapport à 1968, le taux de croissance de la productivité annuelle par tête du Royaume-Uni aura augmenté de 35 %, soit 2,5 % par an, ce qui est peu par rapport à l'Allemagne de l'Ouest : 3,8 %, et à la France : 4,7 %. Pourquoi ce faible chiffre ? Il y a un sur-emploi de main-d'œuvre faiblement productive, beaucoup d'absentéisme (dans le secteur minier, spécialement). Il a été prévu d'élargir l'éventail des emplois dans les zones actuellement déprimées (le Nord du pays) en créant des « Pôles de développement » et les résultats atteints ne sont pas négligeables. D'autre part, la « Selective Employment Tax » pénalise les employeurs qui utilisent une main-d'œuvre nettement excédentaire.

L'investissement annuel moyen de la décennie qui vient de s'écouler pris en tant que pourcentage du P.N.B. reste des plus médiocres : 16 %, contre 24 % en Allemagne de l'Ouest, et 19 % en France.

Quant à la consommation privée, en 1980, la dépense de consommation globale atteindra, comme c'est le cas aujourd'hui, 75 % du P.N.B., par contre la dépense de consommation réelle du ménage britannique moyen s'élèvera de 25 % au-dessus de son niveau actuel : le tableau peut paraître sombre ; une étude, « Family Expenditure Survey », parue en 1966, fournit plus de détails : les gains à attendre dans les 12 années à venir apparaîtront sous la forme d'un accroissement de la consommation des biens dus aux variations de l'acquisition d'un volume plus important d'appareillage électrique et d'achats massifs de voitures personnelles.

III. L'équipement national

Sur 18 millions d'habitations privées, 5 millions datent de la période antérieure à 1885 la même proportion existe pour les écoles et les installations industrielles ; l'idéal serait bien sûr, de remplacer tous ces vestiges du XIX^{e} siècle, il faudrait non seulement construire du neuf, mais aussi transformer les 5 millions d'habitations. En 1980, chaque ménage désireux de se procurer un logement individuel, soit 50 % de la population, devrait en avoir la possibilité. Le gouvernement aura également à remettre en état les bâtiments scolaires, car le nombre des étudiants augmentera d'à peu près 20 % ; or la moitié des écoles primaires ainsi qu'un cinquième des établissements du second degré datent du siècle dernier.

Sylvie GUIRAUD

Le directeur de la publication : G. DUNOD.

Dépôt légal : 4e trimestre 1969. Numéro 6169, Imprimé en France.

Imprimerie Nouvelle, Orléans. — N° 6110.

CONSOMMATION (ANNALES DU C. R. E. D. O. C.)

1965

N° 1. — Quelle est la rentabilité des capitaux investis dans les logements en location ? — Analyse des phénomènes d'induction (Évolution de l'emploi dans le commerce par région entre 1954 et 1962). — Quelques réactions des ménages à l'égard de leur logement. — Un modèle des dépenses médicales. — La consommation en France de 1963 à 1964.

N° 2. — Analyse économique et planification urbaine. — Louer ou acheter son logement. — Réflexions sur le rôle de l'avenir dans ce choix. — Les produits surgelés. — La consommation des boissons de 1960 à 1963. — La fréquentation des colonies de vacances jusqu'en 1964.

N° 3. — Les études d'armature urbaine régionale. — Quelques problèmes posés par la prévision de la demande en services collectifs. — Conditions de logement et insatisfaction des ménages en 1961. — Les dépenses de location de voitures sans chauffeur.

N° 4. — Le Plan, accélérateur de croissance. — L'ajustement de l'offre de viande à la demande. — Étude de la série épargne des ménages (1950-1964).

1966

N° 1. — Recherche et aménagements urbains.

N° 2. — La consommation des Français en 1964. — Étude bibliographique sur l'utilisation des services collectifs. — L'influence des facteurs économiques sur la consommation médicale. — L'influence de la Sécurité Sociale sur les dépenses médicales des exploitants agricoles.

N° 3. — Les conditions du marché du logement et le comportement des ménages. — La consommation pharmaceutique des Français. — Les loisirs aux U.S.A. — Les jeunes ménages et leurs conditions de logement en 1963. — La consommation en France en 1964-1965.

N° 4. — Une méthode pour étudier la solvabilité de la demande de logement. — La loi et les travaux d'Engel. — Le « Federal Reserve Board » et les recherches sur l'épargne.

1967

N° 1. — Une étude économétrique de la demande de viande. — La consommation des Français en 1965. — Intégration des méthodes d'approche psycho-sociologiques à l'étude de l'épargne.

N° 2. — Un indicateur de la morbidité appliqué aux données d'une enquête sur la consommation médicale. — La diffusion des services collectifs : phénomène économique ou social ? — Les travaux de préparation du V^{e} Plan et l'élaboration d'un modèle national de fonctionnement du marché du logement. — Les conditions de vie des familles.

N° 3. — L'épargne des exploitants agricoles. — Structure et équilibre du marché du textile. — Les dépenses touristiques.

1968

N° 1. — Étude critique de méthodes d'enquête. — Étude sur l'offre et la demande de créance.

N° 2. — Théorie et politique de l'épargne. — Un modèle prévisionnel de la demande de logements. — L'évolution de la consommation de viande.

N° 3. — La consommation et la demande de monnaie. — Valeur prédictive des intentions d'achats au niveau du ménage pris individuellement.

N° 4. — Quelques éléments sur le comportement des propriétaires vis-à-vis du prix du logement acheté et de la mise de fonds versée. — Facteurs « irrationnels » de l'offre d'épargne (recherches allemandes).

1969

N° 1. — L'offre de monnaie par les banques commerciales. — L'économie des services de soins médicaux en France. — L'évolution de la consommation de produits laitiers de 1950 à 1966.

N° 2. — L'économie des services de soins médicaux en France. — La formation de l'épargne liquide (l'exemple du Crédit Mutuel). — Consommation individuelle et consommation collective. — Étude sur la demande en logement des ménages.

N° 3. — Les prix de détail en France par rapport aux autres pays de la Communauté. — La consommation des ménages en France et en Hongrie. — Introduction à l'analyse des données.

SOMMAIRE DES PROCHAINS NUMÉROS

Influence des caractéristiques de l'offre sur le recours aux services collectifs. — La consommation en France de 1959 à 1968. — Classification automatique. — Étude des SICAV.

sommaire

CENTRE DE RECHERCHES ET DE DOCUMENTATION SUR LA CONSOMMATION

45, boulevard de la Gare, PARIS-13e

Tél. POR. 97-59

1969 n° 4

octobre décembre